AD________________

GRANT NUMBER DAMD17-96-1-6145

TITLE: Breast Ultrasound: Computer-Aided Diagnosis Approach to Improving Specificity and Decreasing Observer Variability

PRINCIPAL INVESTIGATOR: Jay A. Baker, M.D.

CONTRACTING ORGANIZATION: Duke University Medical Center
Durham, North Carolina 27710

REPORT DATE: January 1998

TYPE OF REPORT: Annual

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
January 1998

3. REPORT TYPE AND DATES COVERED
Annual (1 Jan 97 - 31 Dec 97)

4. TITLE AND SUBTITLE
Breast Ultrasound Computer-Aided Diagnosis Approach to Improving Specificity and Decreasing Observer Variability

5. FUNDING NUMBERS
DAMD17-96-1-6145

6. AUTHOR(S)
Jay A. Baker, M.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Duke University Medical Center
Durham, North Carolina 27710

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The purpose of this grant is to construct an artificial neural network (ANN) to assist radiologists in differentiating benign from malignant solid breast lesions. The ultrasound (US) examinations and mammograms of sixty-four solid breast lesions that subsequently underwent histologic confirmation were evaluated in a blinded manner. Descriptive terms were chosen to characterize the ultrasonographic and mammographic appearance of the lesions from a pre-defined lexicon. In addition, patient age was recorded. These descriptive terms and patient age were used as inputs to train an artificial neural network to differentiate benign from malignant breast masses. An ANN using only seven US descriptive terms as inputs performed better than a similar ANN previously constructed using mammographic descriptive terms and patient medical history as inputs. In addition, neural networks using combinations of inputs including US and mammogram descriptors and patient age all performed well for training an ANN. Further work includes constructing similar ANNs using more training cases, as well as testing these ANNs using prospective data. Significant interobserver variability in radiologists' descriptions and assessment of breast US exams was also demonstrated. The ANN may help decrease the variability in lesion assessment.

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
22

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

__✓__ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

__✓__ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

J. [illegible] MD 2/1/98
PI - Signature Date

Table of Contents

INTRODUCTION:

The primary purpose of diagnostic imaging of the breast is to detect carcinoma of the breast as early as possible. Early detection allows improved prognosis due to treatment at an earlier stage of disease. The traditional techniques of film/screen X-ray mammography and ultrasonography remain the principal modalities for investigating breast disease.

Although film/screen mammography is a successful screening tool due to its ability to detect approximately 90% of breast cancers (i.e., high sensitivity), the radiographic appearance of benign and malignant breast lesions is similar (i.e., low specificity). Because of this overlap in appearance and an overall conservative approach by physicians, only 10-34% of women undergoing breast biopsy actually have breast cancer (i.e., low positive predictive value) [1-7]. The remaining women have benign breast lesions, which would not warrant surgical intervention.

This relatively low positive predictive value (PPV) of breast biopsy raises several problems. First, many women without breast cancer must endure the discomfort, expense, potential complication, change in cosmetic appearance, and anxiety that a breast biopsy can cause. Moreover, the financial burden to society of these procedures is considerable. Therefore, significant improvement in the diagnostic specificity of breast imaging could have substantial impact on reducing the financial, physical, and emotional costs of widespread screening for breast cancer.

Currently, the only widely accepted role of ultrasound (US) in diagnostic breast imaging is the differentiation of simple cysts from solid breast masses[8] [9]. Although a relatively recent study reports that benign and malignant masses of the breast can be differentiated based on grayscale US features [10], this work has not been duplicated and remains controversial[11] [8]. However, because of its low cost, lack of ionizing radiation, and wide availability, US would be an advantageous modality to assist radiologists in distinguishing benign from malignant breast masses. Although US is not useful in screening for breast cancer[12] [13], it is well positioned to assume an important role in assessing masses identified by screening mammography or physical exam.

Unfortunately, ultrasound presently suffers from two limitations. First, *individual* US features are not sufficiently specific to differentiate benign from malignant breast lesions. Although US findings such as hypoechogenicity, decreased through-transmission of the US beam, angular mass margins, and the presence of detectable blood flow on color Doppler imaging all raise the suspicion for breast cancer, no feature *by itself* has proven to have sufficient diagnostic accuracy.

The second limitation to the widespread use of US imaging in differentiating benign and malignant breast masses is that ultrasound is highly operator-dependent. A sonographer must determine the specific location in which to scan as well as define such technical settings as the depth of the focal zone, overall gain, shape of the time-gain compensation curve, and pre- and post-processing algorithms. Therefore, there is considerable opportunity for extensive inter- and intra-observer variability in not only the images obtained but also in the interpretation of those images.

One possible solution to both problems is an artificial neural network (ANN) to assist radiologists in interpreting US images of the breast. An ANN is a form of artificial intelligence analogous to layers of biological neurons. These networks can be trained to "learn" essential information from a set of data[14] [15]. The structure of an ANN is a set of processing units (nodes) arranged in rows. Input nodes are interconnected by simple calculations with an internal layer of hidden nodes and a single output node. Rather than having a fixed algorithmic approach to a classification problem, an ANN is sequentially presented with a set of supervised training cases – input data paired with the correct output. The ANN modifies its behavior ("trains") by adjusting the strength or "weights" of the connections until its own output converges to the known correct output. Once trained, the network can evaluate a new case of input values by applying the weights learned from the data set on which it was trained.

An ANN may be an appropriate tool to assist radiologists in evaluating US images of the breast because of its ability to capture subtle relationships among *multiple* US findings. The ANN may be able to synthesize the information more efficiently than can radiologists alone to improve the diagnostic accuracy of breast US. Rather than evaluating a single US feature, a neural network can determine nonlinear relationships between multiple findings, which, when combined, have the potential to make breast US a more accurate study for diagnosing breast cancer.

An ANN is also well suited to reduce the inter- and intra-observer variability inherent in interpretations of US examination. ANNs are relatively insensitive to minor variations and noise within data [15] [14] [16] [17] allowing a more consistent response despite variability in radiologists' findings. Further, unlike physicians whose threshold for diagnosing breast cancer varies daily, an ANN – given similar inputs on two different occasions – will always provide a consistent diagnostic output.

BODY:

The body of this report will be presented by addressing each work task as outlined in the Statement of Work in the original proposal.

Technical Objective 1:

Develop an artificial neural network (ANN) to predict biopsy outcomes from US findings.

Task 1. Create a database of ultrasound, mammographic, and physical exam findings, as well as medical and family history data for women with solid breast masses and histologically-proven diagnoses.

Methods 1: The first step in constructing an ANN is to collect a set of data (database) with which to build ("train") the neural network. This data set, or training set, consists of inputs with known outputs. As described below, the inputs for this ANN include terms chosen by radiologists to describe mammogram and US images of breast lesions (descriptors). The "known output" is the biopsy result for each case.

The initial proposal for this grant was for each radiologist to prospectively record their choice of descriptive terms for solid breast lesions they found during clinical examinations. All solid lesions undergoing ultrasound examination between August 1, 1995, and continuing through the sixth month of this grant, December 1996 were to be included. Three important difficulties were encountered requiring alteration of this plan. First, given the hectic pace of an active mammography clinic, radiologists did not consistently record their inputs for each case of a solid breast lesion in a prospective manner. In some instances, physicians did not record their inputs until the end of the workday or neglected to record any inputs for a case. Therefore, bias was introduced into the training set because consecutive cases were not available and, often, only the most interesting cases were recorded.

A second unexpected difficulty requiring a change in the system for data collection was the particularly high inter-observer variability discovered between radiologists describing the same images. A study documenting this variability is described in detail below. Although neural networks are relatively insensitive to small fluctuations in inputs once the network is constructed, they may be heavily influenced by considerable variations in the set of data used to *train* the neural network. Therefore, in order to improve the opportunity for constructing a successful ANN, a single reader/radiologist was utilized to determine all the inputs for each case. Cases were therefore evaluated "retrospectively," although still blinded to biopsy results.

Finally, all cases were initially collected using the US machine available in the breast imaging section as of August 1995. This unit (Acoustic Imaging 5200 S, Phoenix, AZ) provided adequate images for routine clinical use. However, in comparison with state-of-the-art US equipment, the images obtained were relatively low in resolution and image contrast. Two, state-of-the-art, high resolution machines (Siemens Sonoline Elegra, Issaquah, WA) unexpectedly became available to the breast imaging section in August 1997. The unit on which all data had been accumulated up to that point was no longer available for clinical use. Therefore, despite the resultant delay in obtaining cases for the training data set, data obtained using the AI US machine were excluded from use in constructing the training database. Including information from before and after August 1997 would have resulted in a possible source of error in constructing the ANN. However, the data obtained prior to August 1997 may be well suited to test the consistency of the ANN when employed by different radiologists using other US machines.

Given these three developments, a new system for collecting data for the training set was employed beginning in August 1997. A single radiologist retrospectively reviews every breast US completed for which biopsy proof is available. This mechanism eliminates inter-observer variability in the data set used to build the neural network. In addition, it eliminates bias in the data set because *all* cases in which US of a solid breast lesion is completed can be reliably included in the study. Finally, all US images used to train the neural network are acquired from the same US machine (i.e., Siemens Sonoline Elegra). This eliminates one potential source of variability in the training set due to significant differences in technology.

The inputs used from each US exam are the radiologist's description of the breast lesion. The terms available to the radiologist for describing the US appearance of breast lesions are those described by Stavros, et al [10]. The terms used by Stavros were chosen for this study because they are well defined and widely available in the literature. A list of these terms is shown in figure 1. In addition to these descriptors for gray-scale images, other information recorded includes the indication for the US exam (i.e., palpable and/or mammographically visible lesion), the size of the lesion, and the presence of blood flow using color and power doppler imaging.

Results 1: The consequence of changing the system of data acquisition is a reduction in the number of cases available for training the ANN at this point in the grant. Sixty-five cases with biopsy proof were available by the end of the first 12 months of the grant. Fortunately, rather than 100 cases with both US findings and biopsy proof being performed each year as anticipated in the grant proposal, between 150 and 180 cases are being performed, in large part due to the improved availability and accuracy of the new US equipment. In addition, the cases obtained prior to August 1997 will be very useful in testing the final ANN, as described above.

Discussion 1: The consequence of changing the system of acquisition of the training data is that fewer – but more useful - cases are available early in the study for constructing the ANN. The data obtained earlier will not be wasted, however. The data obtained by multiple readers will be used at the end of the study, as part of the testing phase of the ANN. The definite advantage of changing the system of data acquisition is that the data used to train the ANN will be free of inter-observer or inter-machine variability, which significantly increases the likelihood of constructing a successful neural network.

Task 2: Build a neural network from the database to predict the presence of breast cancer. Maximize the specificity while maintaining perfect or near perfect sensitivity. Evaluate this computer-aided diagnosis system using "round-robin" techniques.

Methods 2:

ORGANIZATION OF THE NEURAL NETWORK

The ANN for prediction of breast malignancy was constructed as a three layer feed-forward network with a backpropagation training algorithm. The layers consist of an input layer with 7 input nodes, one hidden layer with 4 nodes, and an output layer with one output node. Each input node corresponds to a radiologist's description of an US feature of the lesion.

CASE SELECTION

One hundred-seventy five (175) women had an abnormal ultrasound examination between August 19 and December 23, 1997. Of those with a solid lesion, 65 underwent needle core biopsy, fine needle aspiration, or open excisional biopsy by January 1, 1998. Sixty-four (64) cases were used to construct the ANN to allow for four-way segmentation of the data as described below. Patients ranged in age from 18 to 80 years with an average age of 50 years. At biopsy, 32 (50%) of the lesions were found to be benign while 32 (50%) were malignant.

NETWORK INPUTS

Each US examination was retrospectively evaluated by a single reader/radiologist who was blinded to the biopsy results. The reader described each lesion using the terms – or descriptors – defined by Stavros, et al [10] **(see figure 1)**. One descriptor was selected for each of seven different categories of morphologic characteristics. These categories include (1)mass shape, (2)mass margin, (3)presence of an echogenic pseudocapsule, (4)presence of calcification within the lesion visible by US, (5)acoustic transmission, (6)lesion echogenicity, and (7)lesion echotexture. In addition, features such as the size of the lesion in radial and anti-radial dimensions and the indication for the exam were also recorded. Presence of bloodflow as documented by color and power doppler imaging was also recorded.

TRAINING AND TESTING

The ANN was initially trained to predict the outcome of breast biopsy using only the 7 US features described above as inputs. Each input node consisted of one of the input morphologic features. The optimal number of hidden nodes is difficult to determine. Four hidden nodes were chosen by trial and error because that number provided the best result. The layer of hidden nodes provides an additional level of flexibility to the network to identify connections between the input features and the diagnostic outcome. The network was trained using a backpropagation supervised training algorithm. The outcome (i.e., biopsy result) of each training case was provided to the network along with the case inputs. A benign biopsy result was assigned the value zero while a malignant result was assigned the value one. Output of the neural network ranged continuously from zero to one.

The weights connecting the nodes were initialized to small random values, and the network was trained using a backpropagation algorithm. With this technique the inputs are sequentially applied for each case while the weights of the node connections are iteratively adjusted. The weights converge by minimizing the mean squared difference between the known, correct outcome (benign=0, malignant=1) and the network output (range from 0 to 1). Each subsequent training pair (inputs and biopsy result) was presented to the network and the weights were altered to provide consistent output results. Training continued until the mean squared error was minimized. The information "learned" by the network is stored in the interconnection weights. These weights are not changed once training has been completed.

The network was tested using the "cross validation" technique rather than "round robin" as originally described in the grant. Unlike round-robin testing, cross-validation allows determination of an ROC curve for testing of the ANN. Four-way cross validation was used which entails dividing the 64 cases into four groups of 16 cases. The network is trained using three groups (48 cases) and tested using one group (16 cases). This process is repeated until each group of 16 is used to test the neural network one time.

The outputs of the test cases are recorded as the area under a receiver operating characteristic (ROC) curve (A_z), for which the optimum value is 1.0 and random guessing provides a result of 0.5. The results of the neural network were compared with prior work, which generated ROC areas (A_z) for a similar network that used mammogram descriptors and patient history to predict outcomes of breast biopsy[18]. The positive predictive value (PPV) for the network and radiologists was also compared. More formal statistical analysis of this network will be completed at the conclusion of this grant when a finalized version of the ANN is available.

Results 2: Using only the seven US descriptors described above, the area under the ROC curve (A_z) for the ANN to predict breast biopsy results was calculated. The A_z measured 0.96. The standard deviation of this result was not calculated and will not be calculated until an optimized neural network with sufficient cases is constructed near the conclusion of this grant. A histogram of neural network outputs is shown in figure 2. If a threshold value of 0.20 is chosen, this ANN provides a sensitivity of 100% and a specificity of 69%. This compares with the radiologists' performance of a sensitivity of 100% and a specificity of 50%.

Discussion 2: The results of the early training of this ANN are encouraging. This A_z of 0.96 compares favorably with the A_z of 0.89 determined for a previous ANN that used mammographic descriptors and patient histories as inputs[18]. This result is not entirely surprising because the ANN using mammogram findings as inputs was constructed to evaluate a wider range of breast lesions including calcifications, masses, architectural distortion, and asymmetric densities. In contrast, the US ANN was constructed to evaluate only breast lesions visible by US. A neural network that evaluates only breast lesions visible by US but includes mammographic findings and patient history as inputs may provide further improvement in results. This study is described below in Task 3.

TECHNICAL OBJECTIVE 2:

Evaluate the diagnostic accuracy of the neural network system in a clinical setting.

Task 3: Apply the neural network to approximately 100 cases obtained after the sixth month of the project that are not included in those training cases used to develop the neural network. Determine whether the network generalizes from training cases to new test cases. Test different input features to improve the ability of the network to generalize.

No work has been undertaken to evaluate the diagnostic accuracy of the ANN in a clinical setting. This task requires a completed, optimized neural network to be applied prospectively to actual clinical cases to assess the clinical accuracy, sensitivity and specificity when used in a clinical setting. At least 100 cases will be obtained to train the neural network before this work will be undertaken.

Studies testing different combinations of input features to improve the ability of the network to generalize are listed as follows:

Methods 3:

Prior work [19] has demonstrated that the only piece of medical history that adds useful information to a neural network to predict breast cancer is the patient's age. Therefore, a second ANN was constructed using the seven US descriptors described above and patient age. In addition to these eight input nodes, four hidden nodes were used to construct the ANN. Other information including the indication for the US exam (ie., incidental finding, mammographic abnormality, palpable abnormality, or palpable and mammographic abnormality) have been collected but have not yet been included in preliminary training of an ANN

A third ANN was trained using US features and mammographic features as inputs. In addition to determining US descriptors, the radiologist evaluating each case also evaluated the mammogram for each case. Ten mammographic features were determined for each case. Of these ten features, six were chosen to include in this ANN with expanded inputs. These six features include (1)mammographic appearance of mass margin, (2)mass shape, (3)mass density, (4)associated findings, (5)special cases, and (6)calcification description. These mammogram descriptors use terms defined by the Breast Imaging Reporting and Data System (BI-RADS), a standardized lexicon of morphology terms devised by the American College of Radiology[20]. This third ANN was constructed using the seven US features and the six mammographic features. These 13 inputs were used to construct an ANN with six hidden nodes.

A fourth ANN employed 14 inputs – the 13 imaging findings plus the patient's age – and six hidden nodes to predict the likelihood of breast cancer.

Results 3:

Using the seven US descriptors and patient age to predict the results of breast biopsy, the area under the ROC curve (A_z) for the ANN measured 0.98. A histogram of neural network outputs is shown in figure 3. If a threshold value of 0.35 is chosen, this ANN provides a sensitivity of 97% and a specificity of 90%.

If the seven US features and six mammographic features are used, the area under the ROC curve (A_z) for the ANN to predict breast biopsy measured 0.94. A histogram of neural network outputs is shown in figure 4. If a threshold value of 0.25 is chosen, this ANN provides a sensitivity of 100% and a specificity of 81%. When age is added to the imaging features, the area under the ROC curve (A_z) for the ANN to predict breast biopsy measured 0.96. A histogram of neural network outputs is shown in figure 5. If a threshold value of 0.20 is chosen, this ANN provides a sensitivity of 100% and a specificity of 78%.

Discussion 3:

The results of including inputs such as mammographic features and the patient's age suggest the possibility of improved specificity of the ANN without sacrificing sensitivity. This work is preliminary, however, due to the relatively small number of training cases available. In addition, the threshold values selected must be tested prospectively with additional cases to confirm the accuracy of the networks. The optimum combination of inputs will be sought after 100 to 200 training cases are available.

TECHNICAL OBJECTIVE 3:

Evaluate the usefulness of the ANN in improving observer variability in US examination of breast masses.

Task 4 and 5: Create a database of approximately 100 cases in which three radiologists each independently complete ultrasound examinations of the same solid nodules. Calculate the inter-observer variability of the radiologists' findings of breast US examination of these 100 cases. Use Cohen's kappa statistic to measure observer variability.

Background Task 4 and 5:

Potential exists for considerable inter-observer variability in physicians' interpretation of breast US exams. The instrumentation of US machines requires the user to select several settings including the transducer frequency, depth of view, depth of focal zone, overall gain, time-gain compensation curve, dynamic range, power input, and angle of insonation. Therefore, two sonographers imaging the same lesion will necessarily obtain different images. These may be very slightly different or markedly different, depending on what settings are chosen.

In addition to the issue of obtaining different images, radiologists may interpret the same images in slightly or markedly different ways. Because ANNs are relatively insensitive to subtle variations in input data, it is possible that they may assist physicians in being more consistent in their decisions. In order to determine if improvement is needed and to measure any improvement, the inter-observer variability of radiologists' interpretation of breast US exams must be determined.

Methods Task 4 and 5:

Sixty consecutive US exams of solid breast lesions were obtained between August 19 and October 24, 1997. At least four views of each lesion were obtained; a radial and anti-radial view, both with and without calipers measuring the dimensions of the lesion, were available for each lesion.

Five board-certified radiologists specializing in breast imaging each independently interpreted the exams. No mammograms were provided for comparison. Each radiologist chose a single descriptor from each of seven categories listed in figure 1. The descriptors were evaluated to determine whether the lesion met the Stavros criteria for benignity [10]. In addition, the likelihood of malignancy based on the US images alone was determined by the radiologists who selected one of four levels of suspicion: (1)benign finding; (2)likely benign finding; (3)suspicious finding; (4)highly suggestive of malignancy.

The inter-observer variability of the radiologists' description and assessment of each lesion was determined by Cohen's kappa statistic[21] [22] [23]. Cohen's kappa statistic is a statistical measure designed to assess agreement between two or more observations for categorical or nominal data. This technique determines the proportion of selections for which observers agree and accounts for the possibility of agreements attributable solely to chance. Perfect agreement is indicated by a kappa value of 1.0, whereas a kappa value of 0 indicates the level of agreement expected by chance alone. Although no absolute scale exists, prior reports have suggested the following levels of agreement between observers for the indicated kappa values: ≤ 0.20, slight agreement; 0.21 – 0.40, fair agreement; 0.41 – 0.60, moderate agreement; 0.61 – 0.80, substantial agreement; and 0.81 – 1.00, almost perfect agreement between observers[24].

Results Task 4 and Task 5:

The statistical analysis of agreement between observers for choosing descriptions of sonographic lesion morphology is shown in **table 1.** The greatest agreement was found for determining lesion shape with kappa value 0.80 ± 0.06. The least agreement was found for determining the presence or absence of an echogenic pseudocapsule with kappa value 0.09 ± 0.06. Kappa for determining one of five assessment classifications was 0.30 ± 0.03.

Discussion 4 and 5:

As expected, considerable variability exists in not only choosing terms for describing solid lesions on US images but also for determining the likelihood that such a lesion is malignant. Systems such as the one developed by Stavros, et al. for determining whether a lesion is malignant rely on radiologists to choose specific terms for describing lesions. The considerable variability in choosing descriptor terms demonstrated in this study might make consistent use of such a system difficult. A potential advantage of an ANN is its relative insensitivity to small variations in input data. Future studies will evaluate whether such an advantage exists in this model.

Task 6: Apply the neural network to the data obtained by each of the three observers to determine a computer-aided assessment of the likelihood of breast cancer. Compare the variability and accuracy of the radiologists' assessments with the consistency and accuracy of the predictions of the neural network using the radiologists' findings as inputs.

No comparison between assessments made by radiologists and the ANN have yet been made. The neural network for predicting breast cancer requires additional cases for optimization. Three observers have each evaluated 60 cases, and this data is available for comparing the variability and accuracy of the radiologists' assessments with the accuracy of the neural network once the ANN is optimized.

CONCLUSIONS:

The purpose of this grant is to construct an artificial neural network to help radiologists predict the likelihood of breast cancer from ultrasound images of breast lesions. In order to build this ANN, a set of training cases and testing cases must be collected. Unforeseen deficiencies in the original data collection design and unexpected changes in available instrumentation have resulted in fewer cases available for training an ANN at this point in the grant. However, cases with biopsy proof are becoming available at a faster rate than expected. Changing the system for data collection has eliminated the two major possible sources of error including inter-observer variability in the training data and variability in inputs due to image acquisition using markedly different ultrasound machines. In addition, those cases collected early in this study but not used to train the ANN will be useful as part of the testing set near the end of the study.

A preliminary ANN using seven ultrasound descriptors was trained using 64 cases. The performance of this network was encouraging and was better than a similar network trained on mammographic images and medical history. However, additional training cases are necessary before prospective testing of this ANN can begin. At least 100 to 200 training cases will be collected to optimize the training dataset.

Additional neural networks have also been constructed from the data acquired. These three networks use inputs including: (1) US findings and the patients' age, (2) US findings and descriptions of mammographic findings, and (3) US and mammographic findings and patient age. The addition of age and mammographic features appears to improve the overall specificity of the US neural network. However, additional cases are required to determine whether there is a statistically significant improvement over using US findings alone.

Finally, an inter-observer variability study using five observers and 60 cases demonstrates significant variability in radiologists' description of US images of breast lesions. In addition, there was considerable variability in radiologists' assessment of the likelihood of malignancy for these breast lesions based on the US images alone. Although it is likely that the ANN will decrease this variability, this premise cannot be tested until sufficient training cases are obtained to optimize a neural network to predict the results of breast biopsies.

References

1. Bassett, L.W., *et al.*, *The prevalence of carcinoma in palpable vs impalpable, mammographically detected lesions.* American Journal of Roentgenology, 1991. **157**: p. 21 - 24.
2. Ciatto, S., L. Cataliotti, and V. Distante, *Nonpalpable lesions detected with mammography: review of 512 consecutive cases.* Radiology, 1987. **165**: p. 99-102.
3. Hall, F.M., *Screening mammography: potential problems on the horizon.* New England Journal of Medicine, 1986. **314**: p. 53-55.
4. Kopans, D.B., *The positive predictive value of mammography.* American Journal of Roentgenology, 1992. **158**: p. 521-526.
5. Moskowitz, M., *Screening is not diagnosis.* Radiology, 1979. **133**: p. 265-268.
6. Schwartz, G.F., *et al.*, *Mammographically detected breast cancer: nonpalpable is not a synonym for inconsequential.* Cancer, 1994. **73**: p. 1660-1665.
7. Sickles, E.A., *et al.*, *Medical audit of a rapid-throughput mammography screening practice: methodology and results of 27,114 examinations.* Radiology, 1990. **175**: p. 323-327.
8. Jackson, V.P., *Management of solid breast nodules: what is the role of sonography?* Radiology, 1995. **196**: p. 14-15.
9. Venta, L.A., *et al.*, *Sonographic evaluation of the breast.* Radiographics, 1994. **14**: p. 29-50.
10. Stavros, A.T., *et al.*, *Solid breast nodules: use of sonography to distinguish between benign and malignant lesions.* Radiology, 1995. **196**: p. 123-134.
11. Jackson, V.P., *Sonography of malignant breast disease.* Seminars in Ultrasound, CT, and MR, 1989. **10**(2): p. 119-131.
12. Bassett, L.W. and C. Kimme-Smith, *Breast sonography.* American Journal of Roentgenology, 1991. **156**: p. 449-455.
13. Jokich, P.M., D.L. Monticciolo, and Y.T. Adler, *Breast ultrasonography.* Radiologic Clinics of North America, 1992. **30**(5): p. 993-1009.
14. Boone, J.M., G.W. Gross, and V. Greco-Hunt, *Neural networks in radiologic diagnosis: I. Introduction and illustration.* Investigative Radiology, 1990. **25**: p. 1012-1016.
15. Wasserman, P.D., *Neural Computing: Theory and Practice*. 1990, New York: Van Nostrand Reinhold.
16. Wu, Y., *et al.*, *Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer.* Radiology, 1993. **187**: p. 81-87.
17. Erb, R.J., *Introduction to backpropagation neural network computation.* Pharm Res, 1993. **10**: p. 165-170.
18. Baker, J.A., *et al.*, *Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon.* Radiology, 1995. **196**: p. 817-822.
19. Lo, J.Y., *et al.*, *Effect of patient history findings on predicting breast cancer from mammograms using artificial neural networks.* Academic Radiology, 1998. **submitted February 1998**.

20. Radiology, A.C.o., *Breast imaging - reporting and data system (BI-RADS)*. . 1993, American College of Radiology: Reston, VA.

21. Cohen, J., *A coefficient of agreement for nominal scales.* Educ Psychol Meas, 1960. **20**: p. 161-169.

22. Maclure, M. and W.C. Willett, *Misinterpretation and misuse of the kappa statistic.* American Journal of Epidemiology, 1987. **126**: p. 161-169.

23. Soeken, K.L. and P.A. Prescott, *Issues in the use of kappa to estimate reliability.* Med Care, 1986. **24**: p. 733-741.

24. Landis, J.R. and G.G. Kock, *The measurement of observer agreement for categorical data.* Biometrics, 1977. **33**: p. 159-174.

Figure 1 - Ultrasound Descriptors Used to Construct Artificial Neural Network

Mass Shape

- 0 ellipsoid (wider than tall)
- 1 taller than wide (any part of nodule)

Mass Margin

- 0 well-circumscribed
- 1 mild lobulation (3 or less)
- 2 microlobulations (>3 / each 1-2 mm)
- 3 angular margins
- 4 duct extension (radial extension w/in or around a duct *toward* the nipple)
- 5 branch pattern *(multiple* projections in or around ducts extending *away* from nipple)
- 6 spiculation

Thin, Echogenic Pseudocapsule

- 0 absent
- 1 present

Calcification w/in Nodule on US

- 0 No
- 1 Yes

Acoustic Transmission

- 0 enhanced through transmission
- 1 normal sound transmission
- 2 shadowing/decreased transmission (for any part of mass)

Mass Echogenicity (c/w fat)

- 0 intensely hyperechoic
- 1 isoechoic
- 2 mildly hypoechoic
- 3 markedly hypoechoic (solid)

Echotexture

- 0 homogeneous texture
- 1 heterogeneous texture

Figure 2

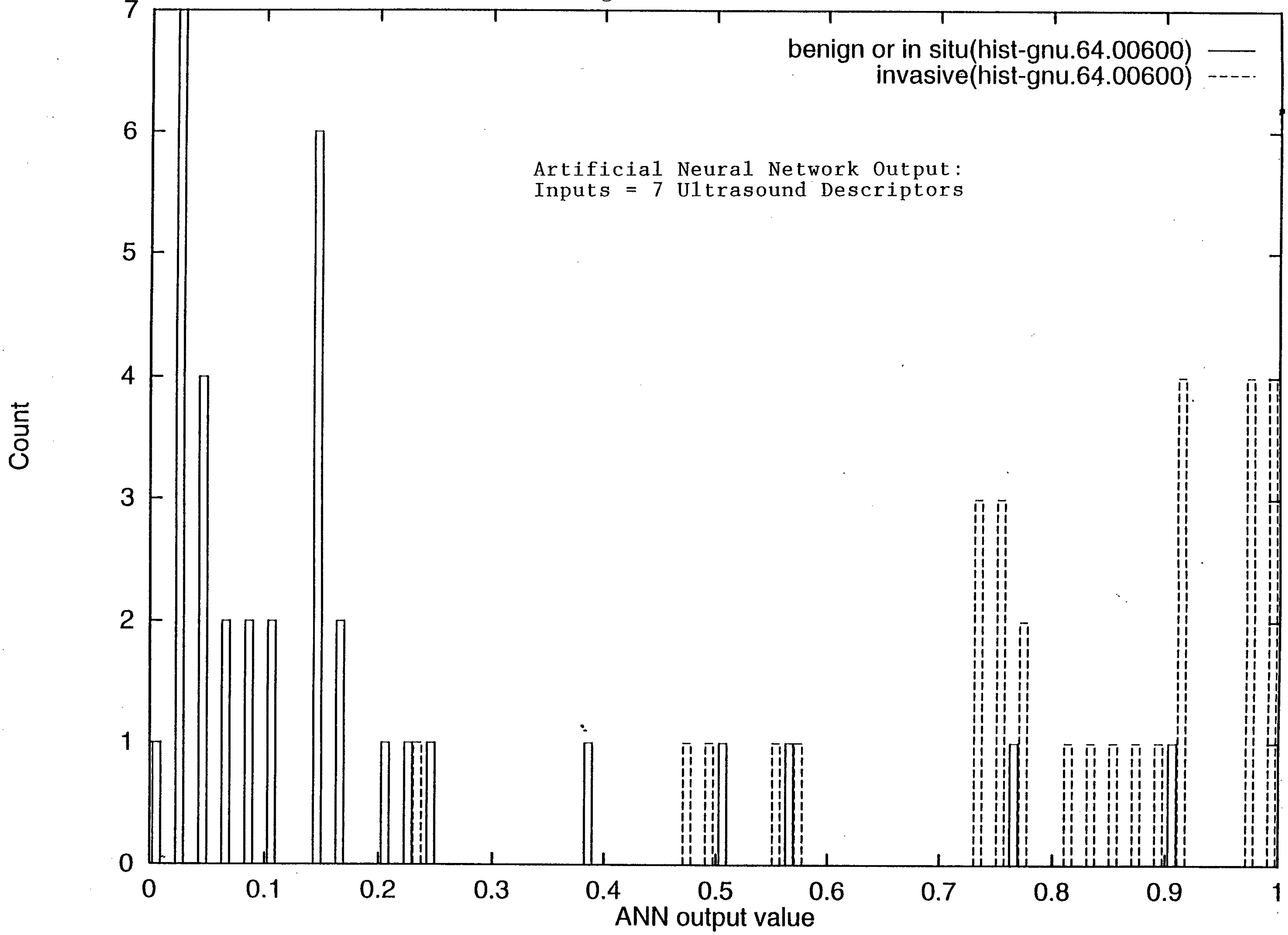

benign or in situ(hist-gnu.64.00600)
invasive(hist-gnu.64.00600)
Artificial Neural Network Output:
Inputs = 7 Ultrasound Descriptors
Count
0
1
2
3
4
5
6
7
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
ANN output value

Figure 3

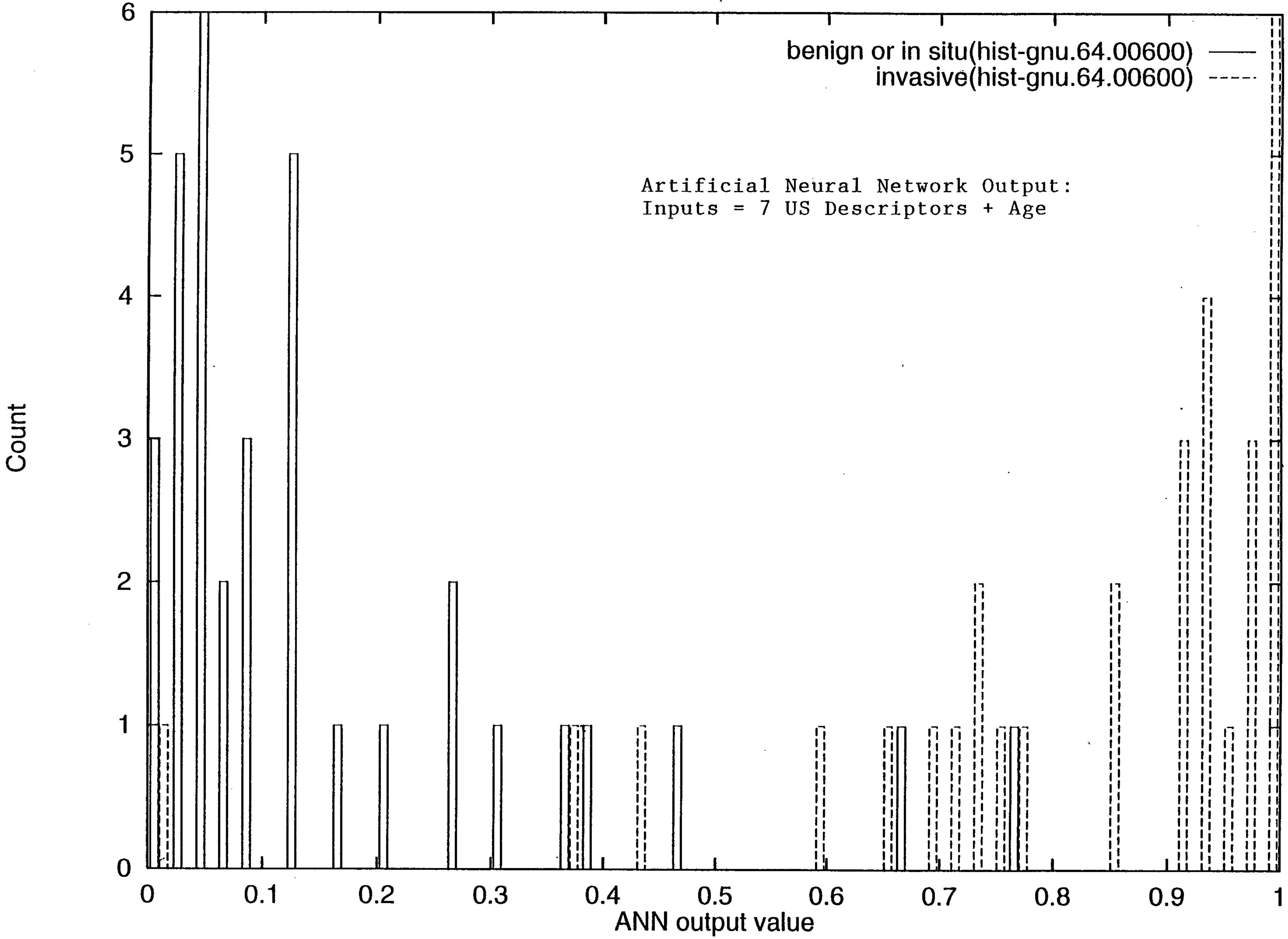
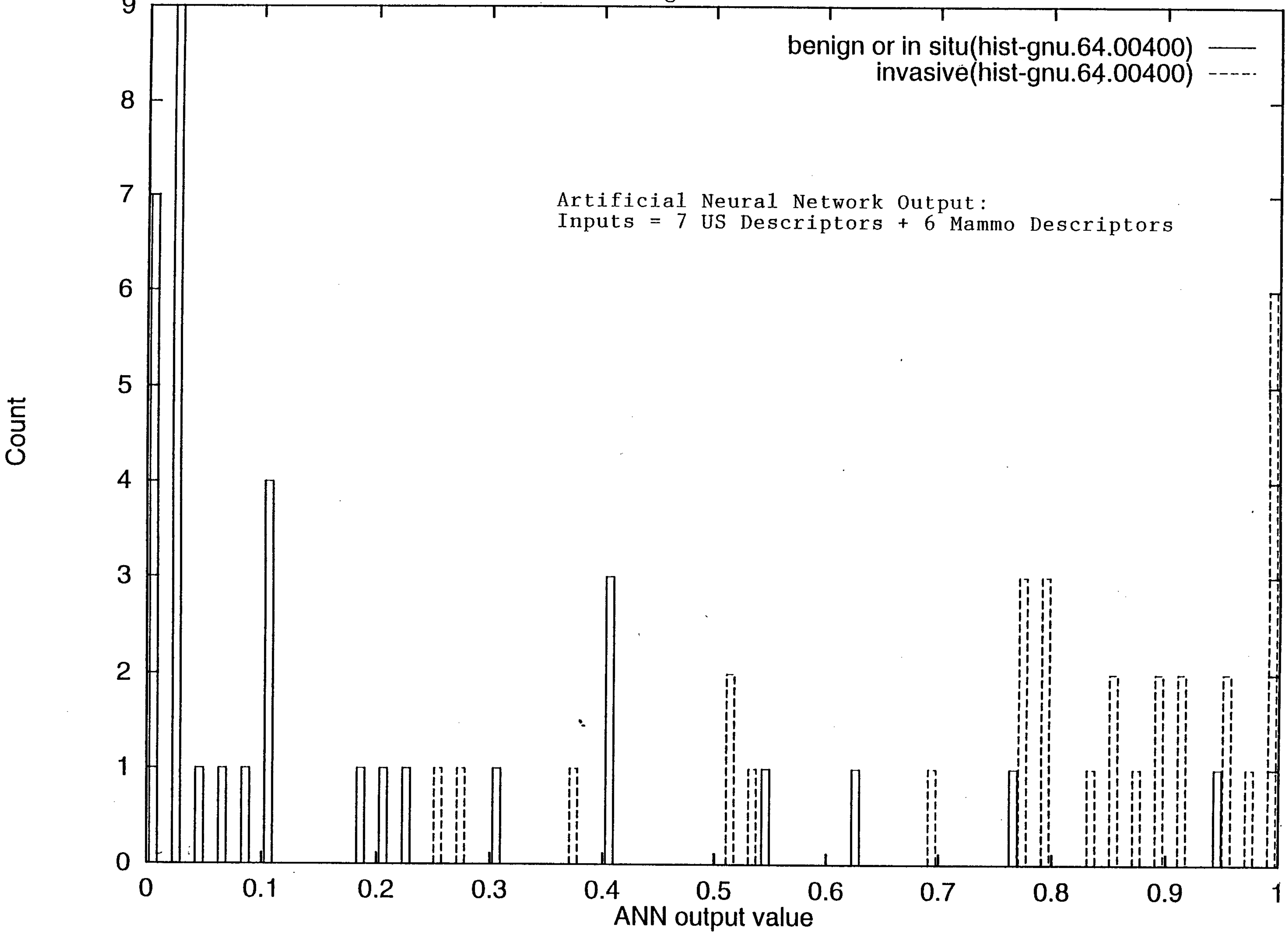

benign or in situ(hist-gnu.64.00600)
invasive(hist-gnu.64.00600)
Artificial Neural Network Output:
Inputs = 7 US Descriptors + Age
Count
0
1
2
3
4
5
6
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
ANN output value

Figure 4


benign or in situ(hist-gnu.64.00400)
invasive(hist-gnu.64.00400)
Artificial Neural Network Output:
Inputs = 7 US Descriptors + 6 Mammo Descriptors
Count
0
1
2
3
4
5
6
7
8
9
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
ANN output value

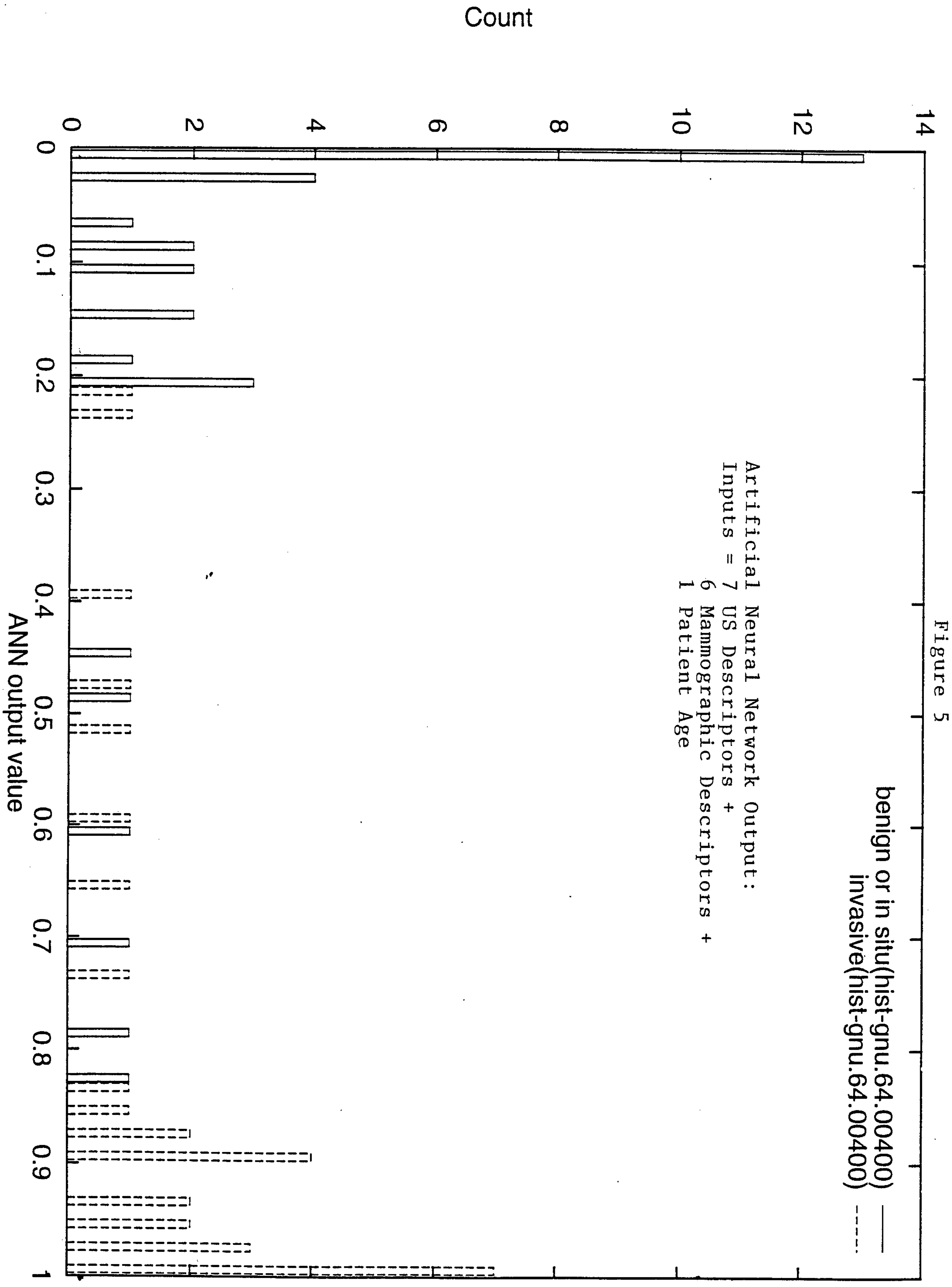


Figure 5

Artificial Neural Network Output:
Inputs = 7 US Descriptors +
6 Mammographic Descriptors +
1 Patient Age

Table 1 - Interobserver Variability of Ultrasound Descriptors for Solid Breast Masses

	Kappa	Std Dev
Mass Shape	0.79	0.06
Mass Margin	0.42	0.04
Pseudocapsule	0.09	0.06
Through transmission	0.55	0.04
Echogenicity	0.4	0.04
Echotexture	0.44	0.06
Likelihood of Malignancy (1-4)	0.3	0.03
Likelihood of Malignancy (1-2) (meets Stavros criteria for benign vs malig)	0.51	0.06

Slight agreement	*< 0.20*
Fair agreement	*0.21 - 0.40*
Moderate agreement	*0.41 - 0.60*
Substantial agreement	*0.61 - 0.80*
Near Perfect agreement	*0.81 -1.00*